Instituto_Nacional de Saúde Doutor Ricardo Jorge
Observações_ Boletim Epidemiológico
2020 número 27
www.insa.pt
artigos breves_ n. 4
_Doenças genéticas e cromossómicas

_Métodos para estudos populacionais de associação genótipo-fenótipo com base em genes candidatos

Methods for candidate gene genotype-phenotype association studies in populations

Susana David[1,2]

suzana.david@insa.min-saude.pt

(1) Departamento de Genética Humana, Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisboa, Portugal

(2) Instituto de Investigação do Medicamento, Faculdade de Farmácia. Universidade de Lisboa, Lisboa, Portugal

_Resumo

Os estudos populacionais de associação genótipo-fenótipo com base em genes candidatos (*Candidate gene Association Studies* – CGAS) conhecem um recente incremento resultante das abordagens baseadas na sequenciação de nova geração. Este artigo resume os princípios gerais do método CGAS que tem vindo a contribuir para a identificação de variantes genéticas causais e para a nossa compreensão da arquitetura genética das doenças humanas.

_Abstract

Population based genotype-phenotype association studies using the candidate gene approach (Candidate Gene Association Studies – CGAS) have seen a recent increase resulting from the use of next generation sequencing methodologies. This article summarizes the general principles of CGAS, that have contributed to the identification of causal genetic variants and to our understanding of the genetic architecture of human diseases.

_Introdução

Os estudos populacionais de associação genótipo-fenótipo com base em genes candidatos (*Candidate gene Association Studies* – CGAS), tais como os estudos de associação do genoma completo (*Whole genome Association Studies* – WGAS), integram o âmbito da epidemiologia genética, uma vaste área de *interface* científica nas Ciências da Saúde. Estes estudos têm vindo a contribuir significativamente para a identificação de variantes genéticas causais e para o aumento da nossa compreensão da arquitetura genética das doenças humanas.

Vários fatores contribuem para o recente incremento da utilização do CGAS em estudos populacionais, sendo que um dos maiores incentivos resulta das abordagens baseadas na sequenciação de nova geração (*Next generation sequencing* – NGS), que se têm revelado eficazes na determinação das sequências de ADN dos genes (regiões codificantes, regiões flanqueadoras, locais de *splicing* e locais regulatórios) e de regiões não codificantes. Este avanço tecnológico foi determinante para a genética humana na encruzilhada da investigação em genética clínica com as abordagens baseadas em estudos populacionais.

_Objetivo

Este artigo resume os princípios gerais do método CGAS aplicado a estudos populacionais como um incentivo para esta abordagem no contexto favorável atual.

_Método CGAS

CGAS versus WGAS

Estudos de associação genótipo-fenótipo recorrem tipicamente a métodos de CGAS ou WGAS. Ambas as abordagens resultaram em importantes contributos para o conjunto de modificadores genéticos hoje conhecidos em medicina humana. Considerando que WGAS é uma estratégia de associação livre de hipóteses *a priori* e não direcionada para a totalidade do genoma, a abordagem CGAS do gene candidato é racional na medida em que o(s) gene(s) e a(s) variante(s) são selecionados com base numa hipótese *a priori* sobre a sua implicação causal na doença [7]. Estes estudos envolvem a genotipagem destes genes candidatos em grupos de indivíduos com fenótipos cuidadosamente definidos, e a análise funcional subsequente de variantes identificadas como estando em associação.

Etapas na realização de CGAS

Existem várias etapas críticas para os CGAS (figura 1). Estas incluem a elaboração do protocolo científico, o projeto de estudo, a definição de fenótipo, a seleção dos genes candidatos, a seleção de variantes candidatas, a análise de haplótipos e de desequilíbrio de ligação (*linkage desiquilibrium* – LD), a estimativa do tamanho de amostra, a significância estatística e valor-*p*, o controle de qualidade dos dados de genotipagem, a análise exploratória dos dados (AED), os testes de associação estatística e a subsequente análise funcional. Estas etapas são seguidamente detalhadas:

- *O protocolo científico*

O protocolo científico, estabelecido antes da realização dos testes de associação, deve especificar a pergunta científica e os objetivos, o desenho do estudo, a hipótese *a priori* e a lógica do estudo, incluindo a definição do(s) fenótipo(s) e a seleção dos genes candidatos e/ou variantes. Deve ainda incluir os testes estatísticos que serão utilizados na análise de associação, a estimativa do tamanho da amostra e a análise do seu poder estatístico.

- *Desenho do estudo*

Quer a abordagem seja a de um estudo de caso-controlo ou de coorte, indivíduos afetados são comparados a indivíduos não afetados. Os estudos de caso-controlo são retrospetivos, sendo menos caros e de duração mais curta do que os estudos de coorte. Por sua vez, estes últimos são geralmente prospetivos e de longa duração. A implicação de variantes genéticas na doença requer que o alelo de risco seja mais frequente, do que seria de esperar apenas pelo acaso, no grupo de indivíduos afetados em comparação com o grupo de indivíduos não afetados (figura 2).

Figura 1: Etapas dos estudos populacionais de associação genótipo-fenótipo com base em genes candidatos.

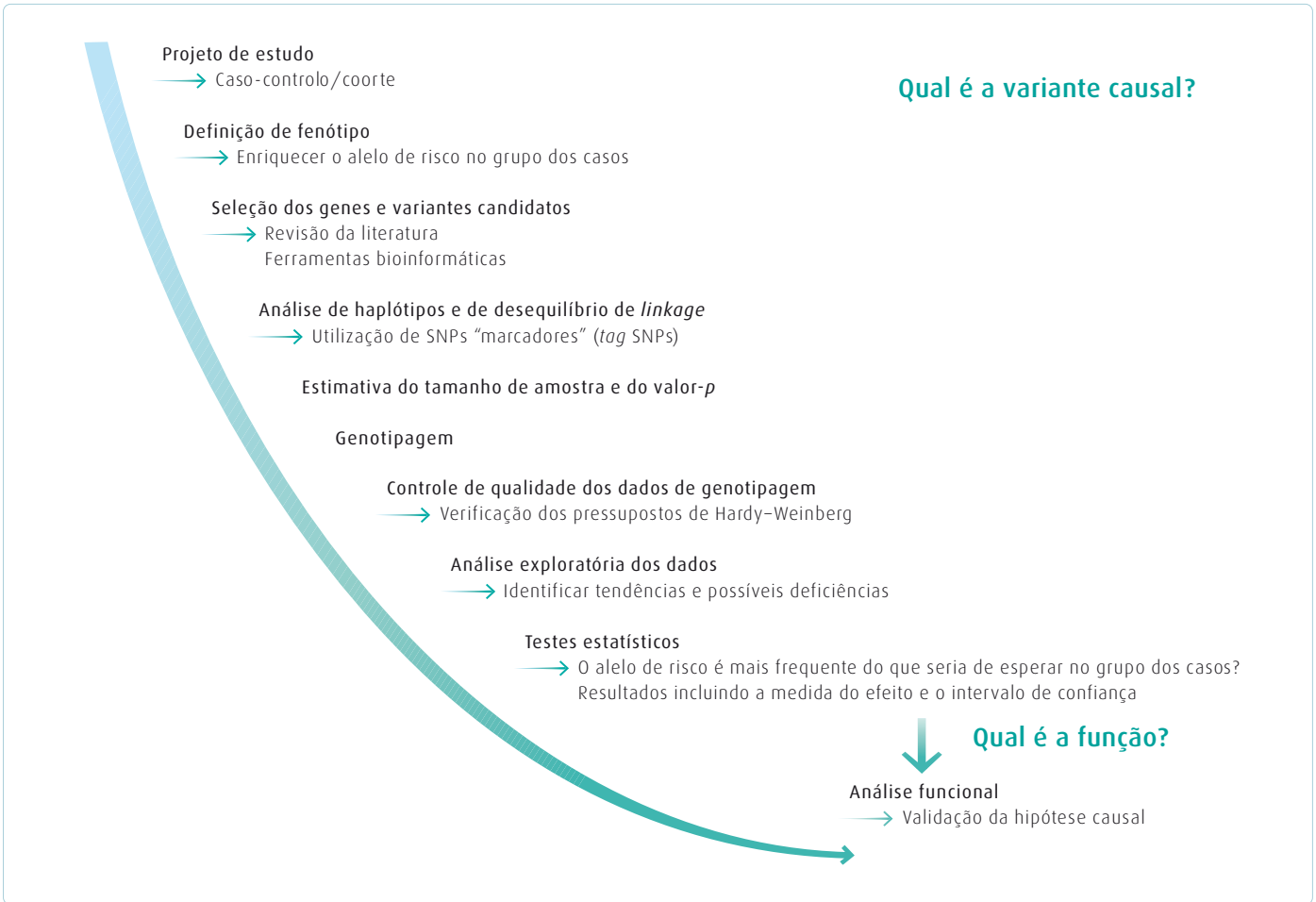
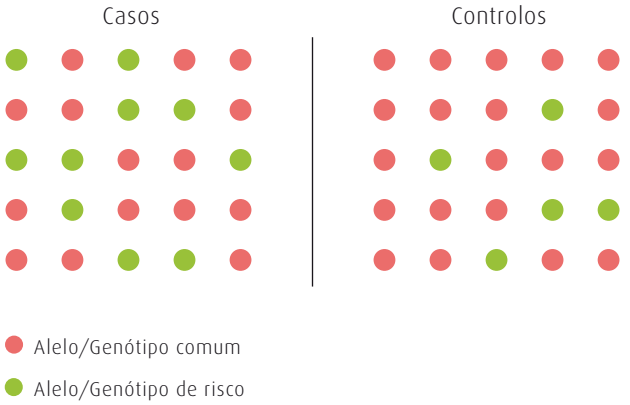


Figura 2: Implicação de variantes genéticas na doença – alelo de risco é mais frequente no grupo de indivíduos afetados (casos) do que no grupo de indivíduos não afetados (controlos).



Definição de fenótipo

O objetivo dos CGAS é detetar a base molecular de uma doença ou ainda das suas demais características biológicas podendo ser traduzidas em termos fenotípicos, como recentemente exemplificámos [12]. A seleção do fenótipo é uma etapa decisiva porque diferentes definições de fenótipo podem levar a resultados diferentes [11]. Da mesma forma, para garantir o sucesso de um estudo poderá não ser suficiente garantir apenas um tamanho de amostra adequado para os testes a realizar. Será também necessário todo o cuidado para reduzir a heterogeneidade nas amostras, resultante, por exemplo, da origem geográfica da população, de fatores ambientais e de diferenças entre grupos de doentes, e das opções de tratamento entre hospitais e serviços hospitalares. Os controlos (grupo de indivíduos não afetados) são selecionados de forma a serem o mais parecidos possível com os casos (grupo de indivíduos afetados), relativamente a todos os atributos relevantes, exceto no caso de não serem afetados ou serem afetados mas de forma diferente. Limitar a definição da doença a um subfenótipo com base em características clínicas particulares também pode ser útil para reduzir a heterogeneidade. Têm sido utilizadas com sucesso várias soluções para enriquecer o alelo de risco no grupo dos casos em estudo, incluindo a seleção de formas mais graves da doença, *i.e.*, "fenótipos extremos", a seleção de formas "proximais" da função do gene (mais "biológicas"), ou de formas relacionadas com o início precoce da doença em que se pode esperar uma maior penetrância da variante (proporção de indivíduos portadores do genótipo que também expressam a característica associada) (figura 2).

Seleção de genes candidatos

Tal como a definição do fenótipo, a seleção criteriosa de genes candidatos é um dos maiores desafios na utilização de CGAS. Como suporte nas estratégias de seleção, várias ferramentas de bioinformática estão disponíveis [8]. Estas incluem ferramentas para pesquisar informações disponíveis na literatura e analisar as vias biológicas participantes, para a priorização dos genes, a análise funcional dos genes e das suas variantes e para explorar correlações entre genótipos e fenótipos. Aqui referimos o Online Mendelian Inheritance in Man (OMIM) ® (*http://www.ncbi.nlm.nih.gov/omim*), que compila relações genótipo – fenótipo conhecidas [2]. Para uma análise abrangente, que fornece informações gerais sobre a estrutura do gene, a expressão, as variantes transcritas, as proteínas codificadas, os elementos reguladores, os polimorfismos de nucleotídeo único (*Single Nucleotide Polymorphisms* – SNPs) e a conservação evolutiva, referimos o Ensembl (*http://www.ensembl.org*) [5] e o UCSC Genome Browser (*http://genome.ucsc.edu/*) [6].

Seleção de variantes genéticas candidatas

A seleção de genes candidatos é frequentemente associada à seleção de variantes genéticas candidatas. O 1000 Genomes Project (*http://www.1000genomes.org/*) [1] e a sua atualizada montagem do genoma humano, GRCh38 (*http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/*) é um recurso importante que fornece aos investigadores frequências de alelos de SNPs num catálogo que abrange variantes de várias populações humanas.

Análise de haplótipos e desequilíbrio de ligação (LD)

O genoma humano contém um número importante de SNPs com tendência para a co-segregação, *i.e.*, serem herdados em conjunto, de modo que é conceptualmente possível selecionar um número limitado de SNPs, ou SNPs "marcadores" (*tag* SNPs), que, na sua totalidade, vão reter a grande parte

da variabilidade genética do conjunto [4,9]. Um haplótipo é um conjunto de alelos em variantes genéticas, num único cromossoma, que estão suficientemente próximos para serem herdados em conjunto. O desequilíbrio de ligação (*linkage disequilibrium* – LD) refere o facto de alelos específicos das variantes genéticas poderem co-ocorrer no mesmo haplótipo numa frequência maior de que a esperada ao acaso [10]. Uma associação genótipo-fenótipo pode ser direta, quando o *locus tag* representa a própria mutação causal do fenótipo de doença em apreço, ou indireta, quando este está próximo e co-segrega com o *locus* da doença. No caso de uma associação indireta, para que a associação entre o alelo marcador e o estado da doença seja observada, os alelos da doença e do *locus tag* devem estar em LD.

Estimativa do tamanho da amostra

No teste estatístico, a análise de poder pode ser utilizada para obter uma estimativa do tamanho mínimo da amostra necessário para detetar uma associação, dado o nível de significância desejado, o tamanho do efeito esperado e o poder estatístico (*i.e.*, a probabilidade de detetar um efeito, se houver um efeito verdadeiro presente para detetar).

Significância estatística e valor-p

O valor-*p* (*p-value*), ou a probabilidade calculada num teste de hipótese, é a probabilidade de obter uma estatística de teste igual ou mais extrema que a estatística observada, assumindo a hipótese nula como verdadeira. O resultado do teste é considerado estatisticamente significativo quando o valor de *p* é inferior a um nível predefinido de significância, ou probabilidade de um erro de tipo I, alfa (geralmente definido como 0,05).

Controlo de qualidade dos dados de genotipagem

A qualidade dos dados da genotipagem é fundamental para o sucesso de um CGAS. A qualidade do DNA é, claramente, um fator importante para garantir a qualidade da genotipagem. No entanto, como os atuais procedimentos de atribuição de SNP são automatizados, podem ser propensos à introdução de genótipos errôneos. Uma das principais advertências sobre esta situação é a não satisfação dos pressupostos de Hardy–Weinberg (Hardy–Weinberg *equilibrium* – HWE). Assim, a verificação do HWE é frequentemente utilizada no sentido de alertar para possíveis problemas na qualidade da genotipagem.

Análise exploratória dos dados

A análise exploratória dos dados (AED) é uma etapa fundamental utilizada para identificar tendências nos dados, bem como possíveis deficiências no conjunto de dados. A AED inclui a avaliação da qualidade dos dados de genotipagem e é útil na seleção de estratégias para aumentar a potência do estudo, por exemplo, reduzindo o número de testes a efetuar. Pode assim ser evitado o problema dos testes múltiplos que ocorre quando muitos testes de hipótese são realizados sobre a mesma amostra, o que pode levar a um aumento de resultados falso positivos [11].

Testes estatísticos

Os métodos estatísticos para testar uma possível associação estão rigorosamente descritos [11]. Assim, a implicação de variantes genéticas na doença requer que o alelo de risco seja mais frequente, do que aquilo que seria de esperar apenas pelo acaso, no grupo de indivíduos afetados (casos) em comparação com o grupo de indivíduos não afetados (controlos). Inversamente, a frequência da variante deverá ser mais baixa no grupo controlo (indivíduos não afetados), bem como na população geral em que o estudo se realiza, assumindo um valor consistente com o tipo de hereditariedade proposto. Em regra, os resultados das análises estatísticas devem ser relatados com a medida do efeito correspondente e o intervalo de confiança apropriado para a medida de efeito subjacente.

Meta-análise

A meta-análise é um elemento de uma revisão sistemática, embora nem todas as revisões sistemáticas incluam uma meta-análise. A meta-análise é um método estatístico de análise de uma grande coleção de resultados de vários estudos científicos independentes. A meta-análise pode ser um recurso, com boa relação custo-benefício, para integrar os resultados de vários estudos e aumentar o seu poder estatístico.

- *Análise funcional*

Quando o critério de significância estatística de uma associação, não é suficiente para implicar uma relação de causalidade, será necessário realizar uma compilação abrangente de evidências de fontes genéticas, bioinformáticas e experimentais para a validação da hipótese causal [7].

_Conclusão

A epidemiologia genética, por meio da seleção de "novos" genes e variantes candidatos, a análise funcional destes e a análise das vias biológicas em que participam, tem vindo a aumentar a nossa compreensão das doenças humanas. A utilização generalizada dos estudos de associação genótipo-fenótipo, CGAS e WGAS, mudou o paradigma dos estudos convencionais de agregação familiar para os estudos populacionais. Houve, assim, uma evolução dos conceitos Mendelianos, de base monogénica com total penetrância e uma relação causal clara, para a integração de distúrbios complexos resultantes de componentes poligénicos ou monogênico geneticamente heterogêneos mas fisiologicamente homogéneos, de baixa penetrância [3]. Espera-se que num futuro próximo estes desenvolvimentos possam ainda contribuir para novas intervenções em saúde, incluindo no diagnóstico, na previsão do risco de doença, na prevenção, na tomada de decisões terapêuticas, na avaliação da resposta terapêutica e no prognóstico, assim como no cumprimento das metas de saúde pública (iniciativa da OMS Human Genomics in Global Health, *https://www.who.int/genomics/about/en/*).

Referências bibliográficas:

(1) 1000 Genomes Project Consortium; Abecasis GR, Auton A, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491 (7422):56-65. https://doi.org/10.1038/nature11632

(2) Amberger J, Bocchini C, Hamosh A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). Hum Mutat. 2011;32(5):564-67. https://doi.org/10.1002/humu.21466

(3) Casanova JL, Abel L. The human genetic determinism of life-threatening infectious diseases: genetic heterogeneity and physiological homogeneity?. Hum Genet. 2020;139(6-7):681-94. https://doi.org/10.1007/s00439-020-02184-w

(4) Chapman JM, Cooper JD, Todd JA, et al. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered. 2003;56(1-3):18-31. https://doi.org/10.1159/000073729

(5) Hubbard TJP, Aken BL, Ayling S, et al. Ensembl 2009. Nucleic Acids Nucleic Acids Res. 2009;37(suppl 1):D690–D697. https://doi.org/10.1093/nar/gkn828

(6) Kuhn RM, Karolchik D, Zweig S, et al. The UCSC Genome Browser Database: update 2009. Nucleic Acids Res. 2009;37(suppl 1):D755–D761. https://doi.org/10.1093/nar/gkn875

(7) MacArthur DG, Manolio TA, Dimmock DP, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508(7497):469-476. https://doi.org/10.1038/nature13127

(8) Patnala R, Clements J, Batra J. Candidate gene association studies: a comprehensive guide to useful in silico tools. BMC Genet. 2013;14:39. https://doi.org/10.1186/1471-2156-14-39

(9) Stram DO. Tag SNP selection for association studies. Genet Epidemiol. 2004;27 (4):365-74. https://doi.org/10.1002/gepi.20028

(10) Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet. 2003;4(8):587-597. https://doi.org/10.1038/nrg1123

(11) Ziegler A, König IR, Pahlke F. A Statistical Approach to Genetic Epidemiology: Concepts and Applications. 2nd ed. Weinheim : Wiley-VCH, 2010.

(12) David S, Aguiar P, Antunes L, et al. Variants in the non-coding region of the TLR2 gene associated with infectious subphenotypes in pediatric sickle cell anemia. Immunogenetics. 2018;70(1):37-51. https://doi.org/10.1007/s00251-017-1013-7